STT 2100-6115
Algorithme de Hasting-Metropolis

Simon Fontaine Nonvikan Karl-Augustt Alahassa

Université de Montréal – Département de Mathématiques et de Statistique

3 avril 2017

Introduction à l'algorithme Hasting-Metropolis

Exemple

Supposons que l'on désire estimer la moyenne d'une distribution π de la forme :

$$\pi(x) = Cw(x)$$

où C est la constante de normalisation, et w une composante intégrable. C'est souvent le cas en statistique bayésienne, lorsque la constante de normalisation est une quantité difficile à calculer [Brooks, 1998].

L'esprit de l'algorithme Hasting-Metropolis est de trouver une chaîne de Markov dont la probabilité de transition satisfait la condition (*detailed balance equation*) :

$$\pi(x)p(x,y) = \pi(y)p(y,x) \tag{1}$$

où $p(x,y)$ est la probabilité de transition de l'état x à l'état y.

Introduction à l'algorithme Hasting-Metropolis

- Si une chaîne (X_t) de Markov est irréductible est apériodique, π sera l'unique distribution stationnaire si la condition (1) est satisfaite. Cette condition est qualifiée de *réversibilité*.
- Intuitivement, cela se comprend : si la chaîne a plusieurs composantes, il est normal qu'elle admette plus d'une distribution stationnaire à la fois ! Si de plus, la chaîne est réversible (i.e condition 1, équivalente aussi à $(X_0, X_1, ..., X_t) \stackrel{d}{=} (X_t, X_{t-1}, ..., X_0)$), la distribution stationnaire est unique, quelque soit le sens dans lequel on parcours la chaîne.
- Lorsque la chaîne de Markov est ergodique (admet une distribution stationnaire), elle y converge à une vitesse exponentielle [Arora, 2013].

Introduction à l'algorithme Hasting-Metropolis

En statistique, l'algorithme MH est une méthode Monte Carlo (MCMC) par chaîne de Markov pour obtenir une séquence d'échantillons aléatoires à partir d'une distribution de probabilité pour laquelle l'échantillonnage direct est difficile.

Hasting-Metropolis

- Considérer des valeurs initiales, $X^{(0)}$, et itérer pout t = 1, 2, 3,...
- Générer $X \sim q(\cdot|X^{(t-1)})$, on considère $q(\cdot|X^{(t-1)})$ comme une densitè instrumentale.
- Calculer la quantité :

$$\alpha(X|X^{(t-1)}) = min\Big\{1, \frac{f(X)q(X^{(t-1)}|X)}{f(X^{(t-1)})q(X|X^{(t-1)})}\Big\}$$

Avec probabilité $\alpha(X|X^{(t-1)})$ prendre $X^{(t)} = X$, ou bien prendre $X^{(t)} = X^{(t-1)}$.

Hasting-Metropolis

- Si $\alpha(X|X^{(t-1)}) < 1$, on tire à pile ou face avec probabilité égale à $\alpha(X|X^{(t-1)})$.
  - Si on gagne, on pose $X^{(t)} = X$,
  - Si on perd, on pose $X^{(t)} = X^{(t-1)}$.
- D'une certaine manière, cet algorithme interpole entre mouvement determinisme (ici, rester sur place), et mouvement stochastique (se deplacer aleatoirement). C'est l'un des points remarquables de cet algorithme.

Independence-Metropolis-Hastings Algorithm (IMHA)

La version (IMHA) est une version *importance sampling* où la densité instrumentale est une densité fixe, g. Dans le cas d'une marche indépendante, la distribution instrumentale doit satisfaire une inégalité de la forme :

$$g(z) \geq \epsilon f(z)$$

L'objectif de cette contrainte est de s'assurer que la convergence de l'échantillonneur à la distribution *stationnaire* est uniformément ergodique. C'est un des résultats principaux de Mengersen and Tweedie (1996).
La probabilité d'acceptation est donc :

$$r(x, y) = min\left\{\frac{f(y)g(x)}{f(x)g(y)}, 1\right\}$$

Il est souhaitable que g soit une bonne approximation de f.

Independence-Metropolis-Hastings Algorithm (IMHA)

Exemple

Cet exemple est celui de [Robert et Casella, 2004] au numéro 7.18 (pp.293-294) dans le livre Monte Carlo Statistical Methods. On considére une distribution Z de densité *inverse gaussienne* :

$$f(z) \propto z^{-3/2} \exp\{-\theta_1 z - \frac{\theta_2}{z} + 2\sqrt{\theta_1 \theta_2} + Log(2\theta_2)\}, z > 0$$

avec $\theta_1, \theta_2 > 0$.

Independence-Metropolis-Hastings Algorithm (IMHA)

Exemple

D'après Brown, 1986, Chapter 2, ou encore Robert, 2001, lemme 3.3.7 : $E[(Z.\frac{1}{Z})] = \left(\sqrt{\frac{\theta_1}{\theta_2}}, \sqrt{\frac{\theta_1}{\theta_2}} + \frac{1}{2\theta_2}\right)$.
Si l'on souhaite générer des valeurs de Z pour θ_1 et θ_2 fixés, le choix d'une densité instrumentale pourrait être porté sur une Gamma $\gamma(\alpha, \beta)$ de densité g, de telle sorte que l'on ait $\alpha = \beta\sqrt{\frac{\theta_1}{\theta_2}}$.
Le rapport :

$$\frac{f(x)}{g(x)} \propto x^{-\alpha-1/2} \exp((\beta - \theta_1)x - \frac{\theta_2}{x})$$

montre que f/g est borné pour $\beta < \theta_1$.

Algorithme Hasting-Metropolis de type marche aléatoire

C'est une particuliarité de l'algorithme de départ

- Si q est symétrique (gaussienne par exemple), le rapport $q(X^{(t-1)}|X)/q(X|X^{(t-1)})$ vaut 1, et l'algorithme s'appelle *Metropolis*.
- La seule différence entre Metropolis et Metropolis-Hastings est que la première utilise une distribution symétrique, et la seconde fut introduite par [Hastings, 1970].
- Dans le type marche aléatoire, la densité q est re-centrée après chaque étape à la valeur générée en dernier lieu par la chaîne. Généralement, la distribution est gaussienne (on prend donc $\mu = X^{(t-1)}$, et $\sigma =$ stepwise), auquel cas la symétrie est satisfaite !
- Si $\mu =$ *constante*, on rejoint le HM de type *indépendant*.

NB: Si σ trop grand, on a des points localement persistants au niveaux de la chaîne, si σ trop petit, le support de π n'est pas couvert (Voir notes de David Draper, 2005, page 146).

Exemple – Simulation d'une normale bivariée

On souhaite simuler des observations selon $\mathbf{X} \sim \mathcal{N}_2(\mu, \Sigma)$, où $\mu = (\mu_1, \mu_2)^\top$ et $\Sigma = (\sigma_i j)_{i,j=1}^2$.

$$f(\mathbf{x}|\mu, \Sigma) \propto \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu)\right\}$$

- Par décomposition de Cholesky, il est possible de générer de telles observations sans recourir à des méthodes MCMC :
  - On simule $\mathbf{Z} \sim \mathcal{N}_2(0, I_2)$ composante par composante en inversant la fonction de répartition d'une normale standard;
  - On calcule $\mathbf{X} = \mu + S^\top \mathbf{Z}$, où $\Sigma = S^\top S$.
- On pourra donc comparer nos simulations à une génération parfaite.

Exemple – Simulation d'une normale bivariée

Supposons que la chaîne de Markov est présentement en $\mathbf{x}$. La génération du candidat $\mathbf{y}$ peut se faire de plusieures manières.

Marche aléatoire

- On considère des candidats générés selon $\mathbf{y} = \mathbf{x} + \mathbf{z}$, où $\mathbf{z}$ provient d'une distribution centrée en 0 et facile à simuler;
- On choisit une densité symétrique :
  - Uniforme $Z_i \sim \mathcal{U}(-\delta_i, \delta_i)$, $i = 1, 2$;
  - Gaussienne $Z_i \sim \mathcal{N}(0, \tau_i^2)$, $i = 1, 2$
- On doit spécifier les paramètres de la distribution pour contrôler la probabilité d'acceptation

$$\alpha(\mathbf{x}, \mathbf{y}) = \min\left(\frac{f(\mathbf{y})}{f(\mathbf{x})}, 1\right)$$

Exemple – Simulation d'une normale bivariée

Marche aléatoire Gausienne : résultat

[Roberts et al., 1997] :

- Pour une marche aléatoire gaussienne et densité cible aussi gaussienne,
- En dimension 1, la probabilité d'acceptation idéale est de 0,45;
- En dimension > 1, la probabilité idéale est de 0,23.

Exemple – Simulation d'une normale bivariée

M.-H. Indépendant

- On doit spécifier complètement la distribution $q(\mathbf{y}) = q(\mathbf{y}|\mathbf{x})$;
- On veut que cette densité aie le même support que $\mathbf{X}$ afin de visiter toutes les régions intéressantes;
- On veut aussi que la distribution approxime bien la distribution de $\mathbf{X}$ afin de contrôler la probabilité d'acceptation donnée par

$$\alpha(\mathbf{x}, \mathbf{y}) = \min\left(\frac{f(\mathbf{y})q(\mathbf{x})}{f(\mathbf{x})q(\mathbf{y})}, 1\right).$$

Exemple – Simulation d'une normale bivariée

M.-H. Indépendant : Loi Gaussienne

- Gaussienne indépendante : on considère des candidats générés selon $Y_i \sim \mathcal{N}_2(0, S)$;

$$q(\mathbf{y}) \propto \exp -\frac{1}{2}\mathbf{y}^\top S^{-1}\mathbf{y};$$

- Le support de cette densité est bien $\mathbb{R}^2$;
- On doit choisir les paramètres de S pour approximer le mieux la densité $f(.)$;
- (Notons que l'on triche en choisissant une telle distribution instrumentale!)

Exemple – Simulation d'une normale bivariée

Application numérique

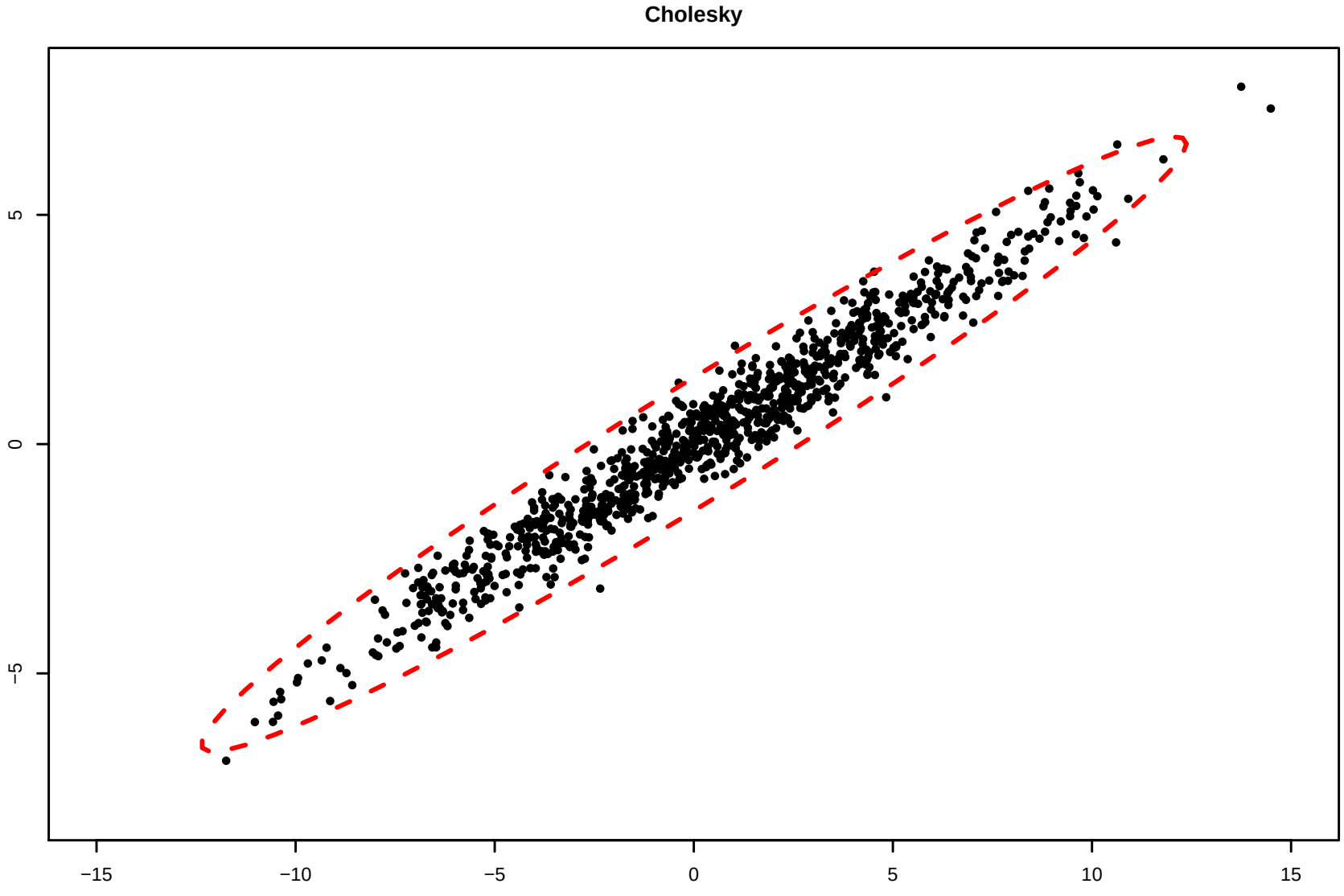
- Distribution cible :

$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} 17 & 9 \\ 9 & 5 \end{pmatrix}.$$

- On considère différentes formes et tailles de densités pour des marches aléatoires uniformes et gaussiennes et pour l'échantillonneur indépendant gaussien.

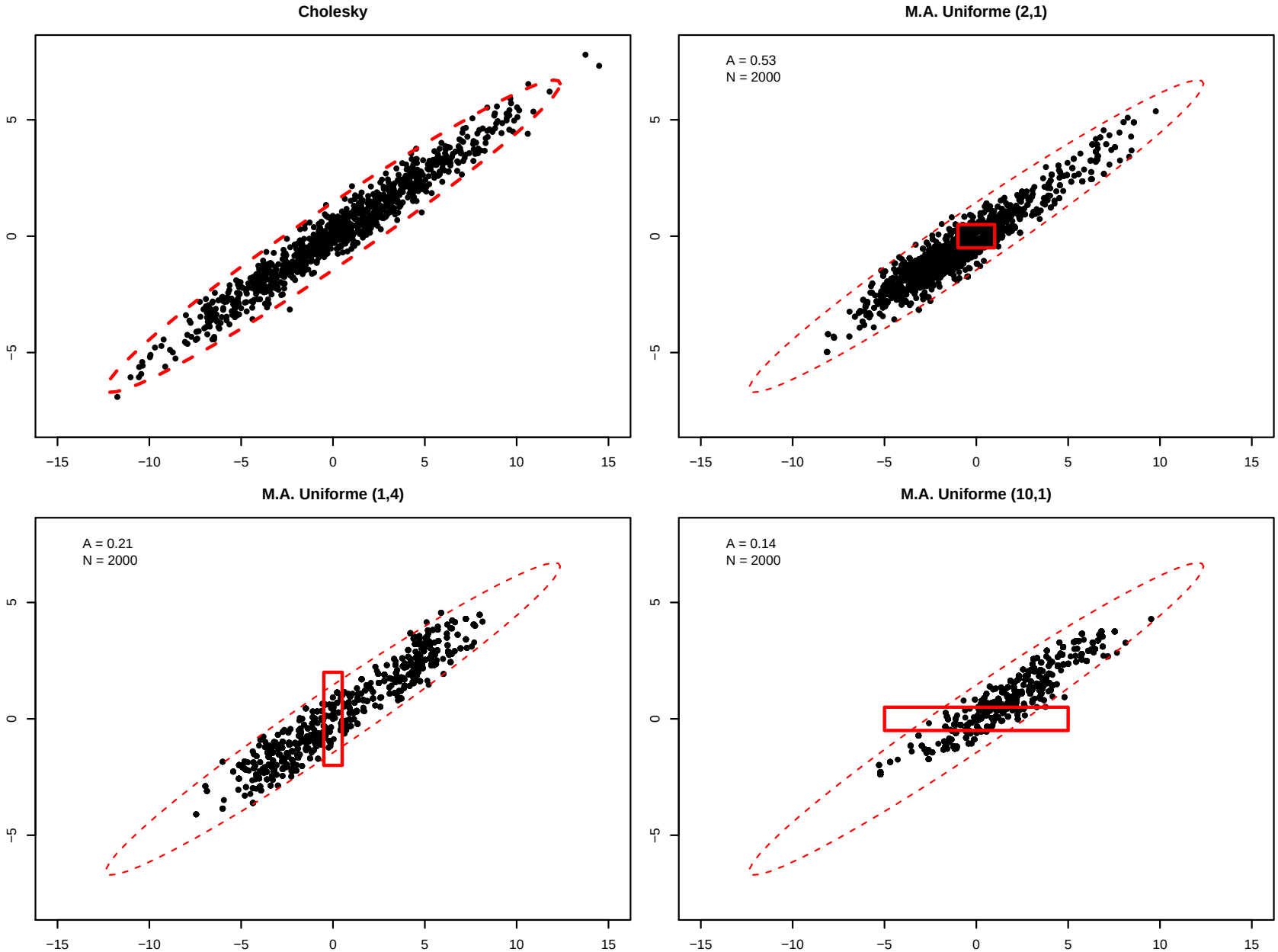
Exemple – Simulation d'une normale bivariée

Échantillonage de Cholesky (N=1000)



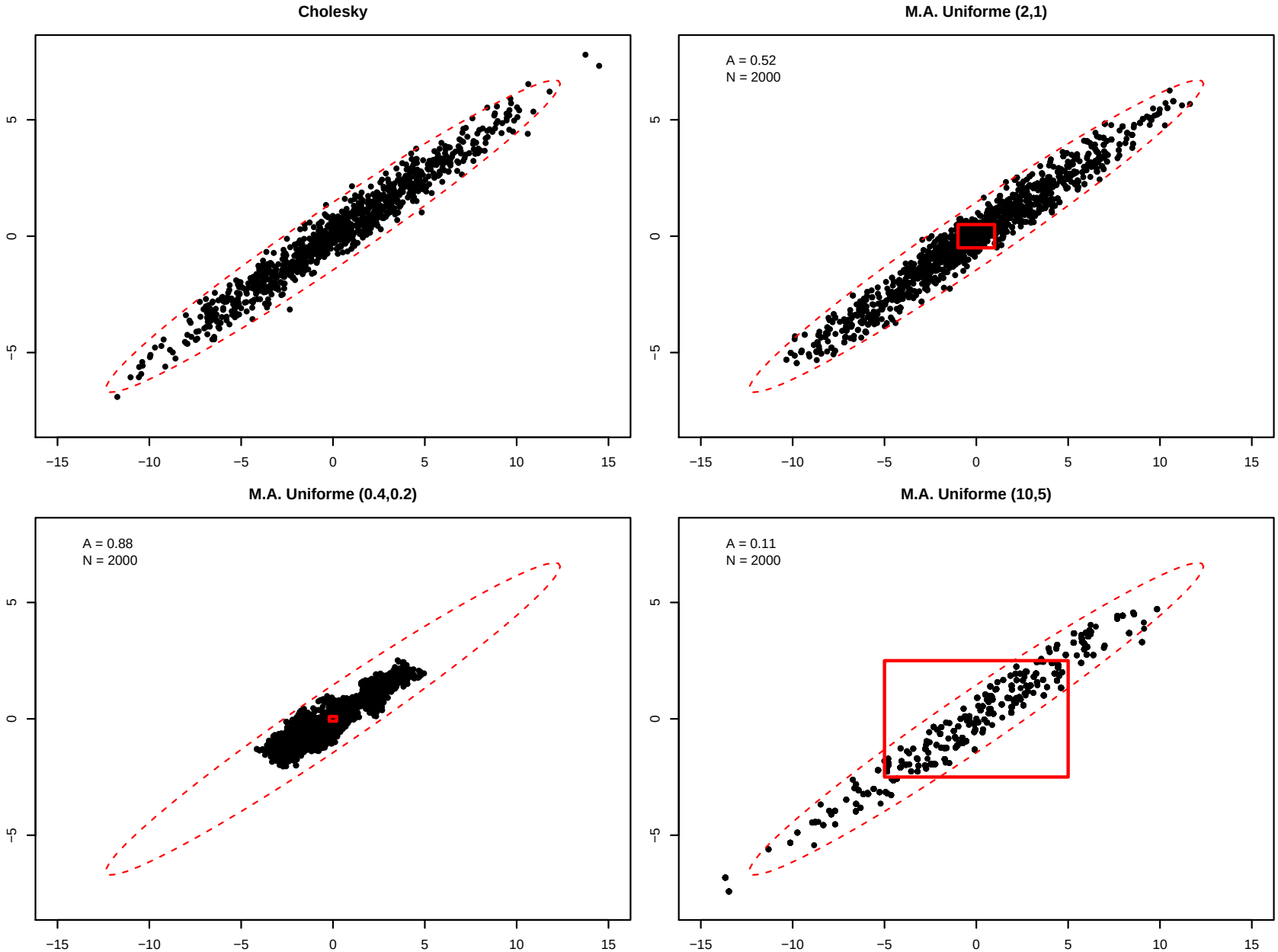
Exemple – Simulation d'une normale bivariée

Marche aléatoire Uniforme : choix de la forme



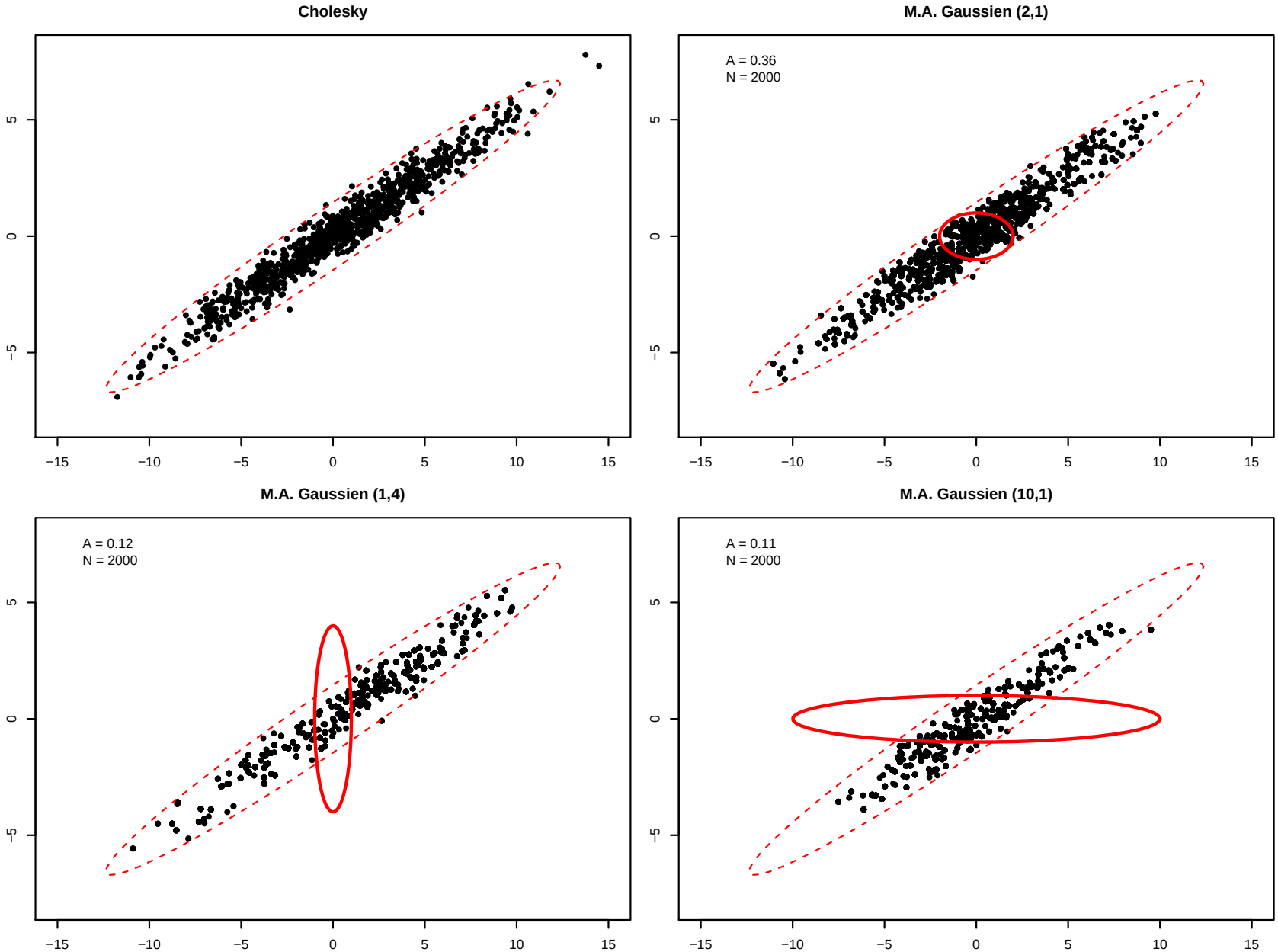
Exemple – Simulation d'une normale bivariée

Marche aléatoire Uniforme : choix de la taille



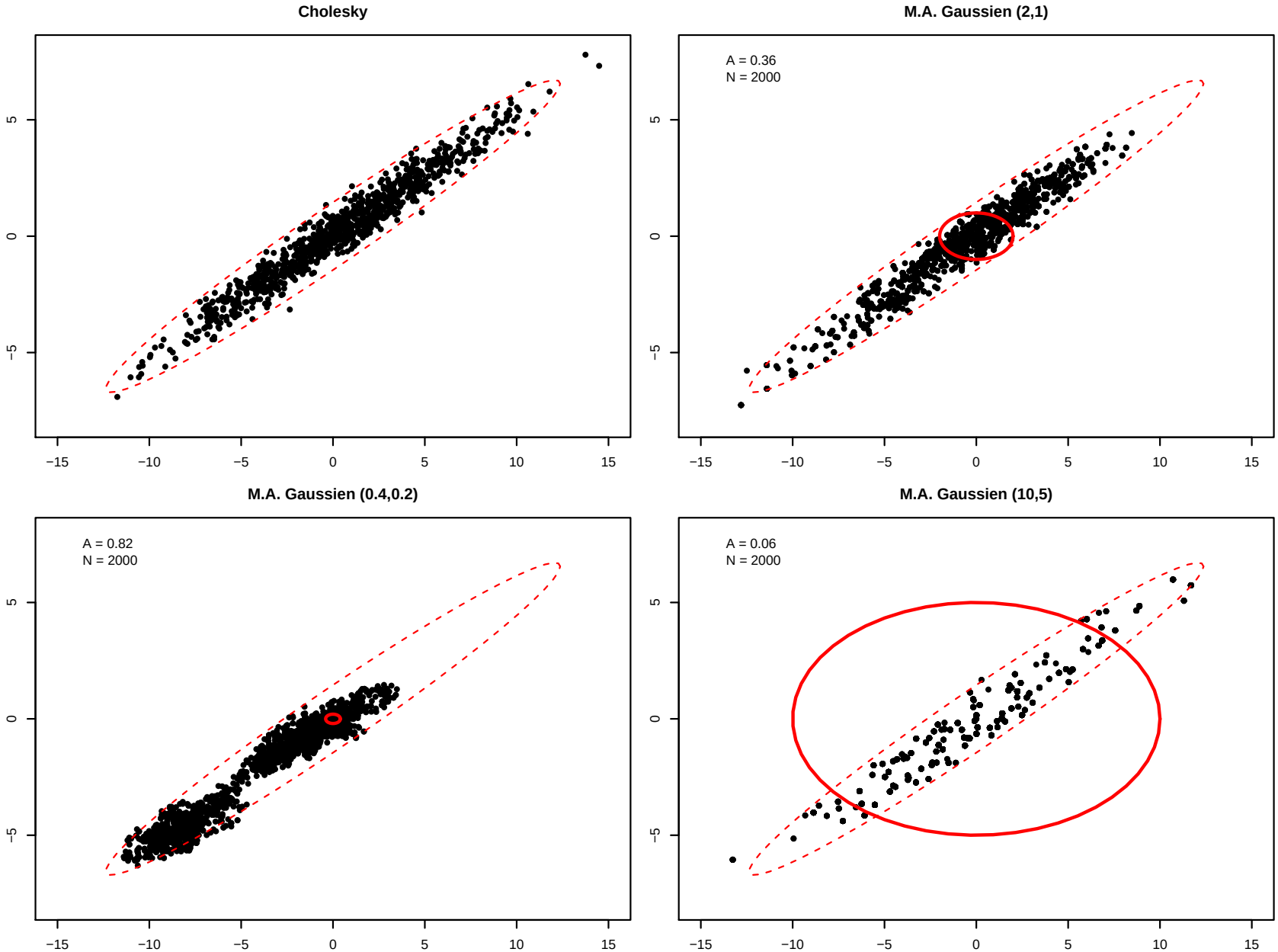
Exemple – Simulation d'une normale bivariée

Marche aléatoire Gaussienne : choix de la forme



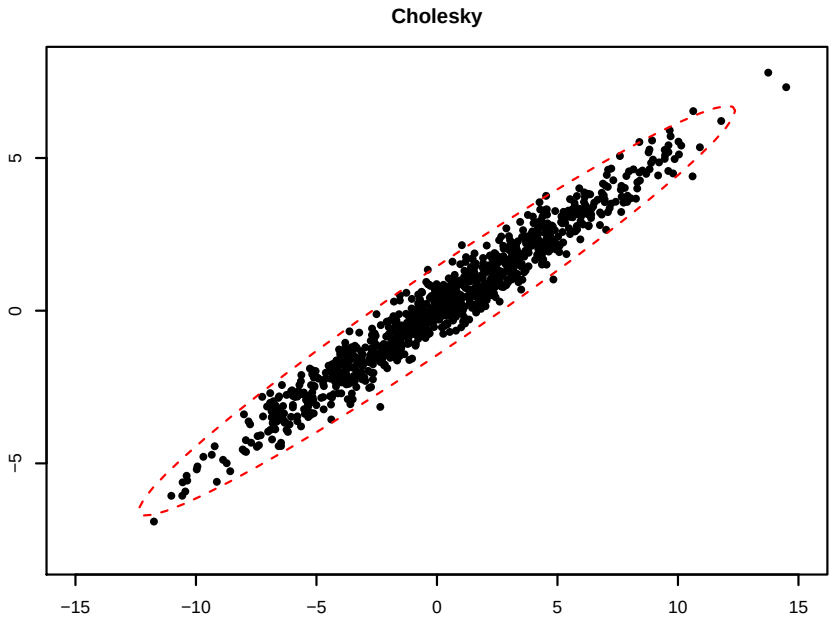
Exemple – Simulation d'une normale bivariée

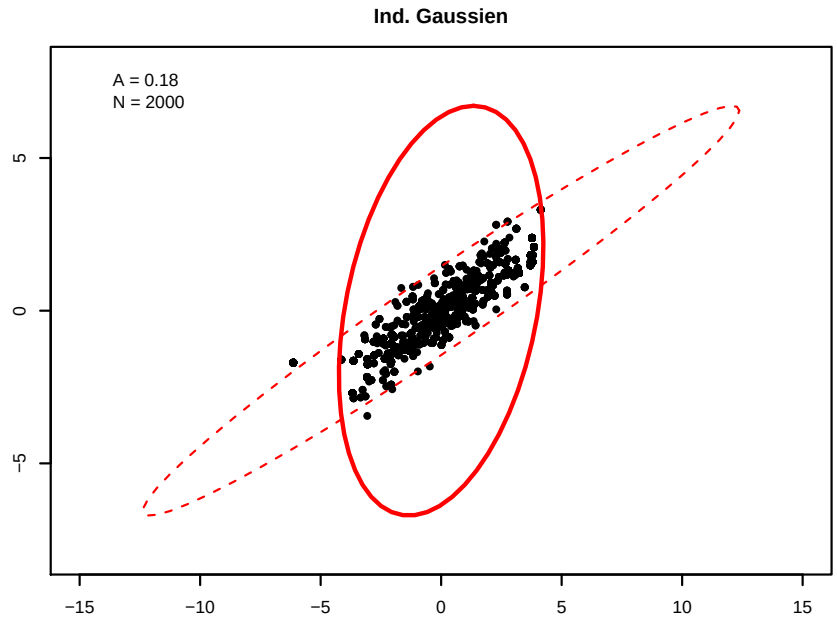
Marche aléatoire Gaussienne : choix de la taille

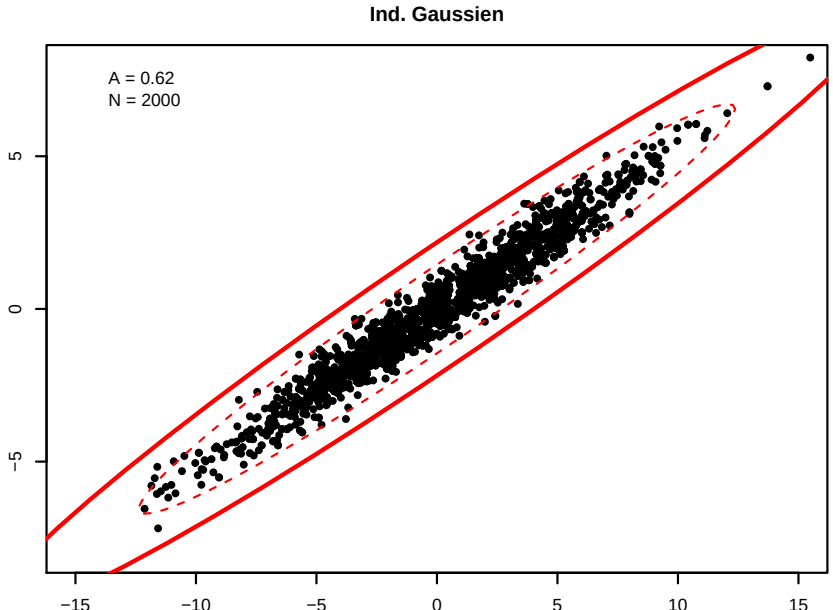


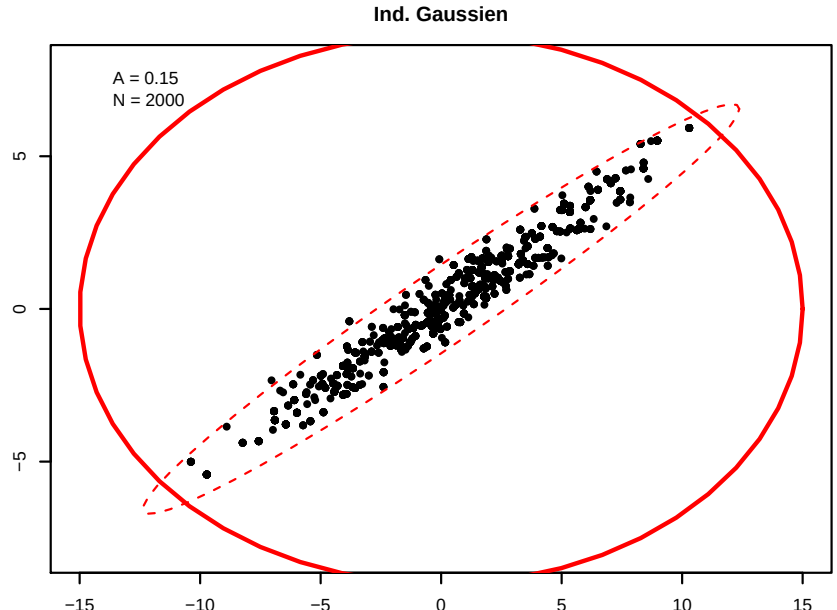
Exemple – Simulation d'une normale bivariée

Inépendante Gaussienne : choix de la forme









Exemple – Simulation d'une loi a posteriori

Justification

- En calcul Bayesien, il est simple de trouver une loi a posteriori, mais le calcul des constantes d'intégrations peut être plus difficile.
- On doit donc recourir à des simulations pour obtenir une approximation de la distribution.

Exemple – Simulation d'une loi a posteriori

Mise en situation

On considère le modèle suivant de mélange de deux lois Poisson à moyenne connue :

$$X_i|\xi \overset{iid}{\sim} \begin{cases} \text{Pos}(\lambda_1, & \text{avec pr. } \theta, \\ \text{Pos}(\lambda_2), & \text{avec pr. } 1-\theta, \end{cases} \qquad i = 1, \ldots, n;$$

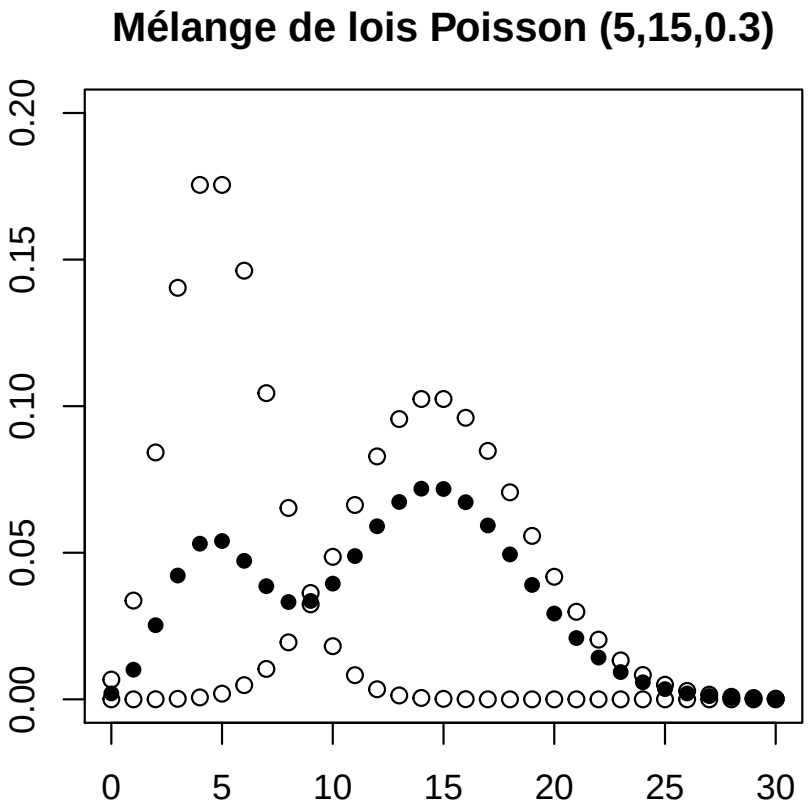
$$\theta \sim \text{Bêta}(\alpha, \beta).$$

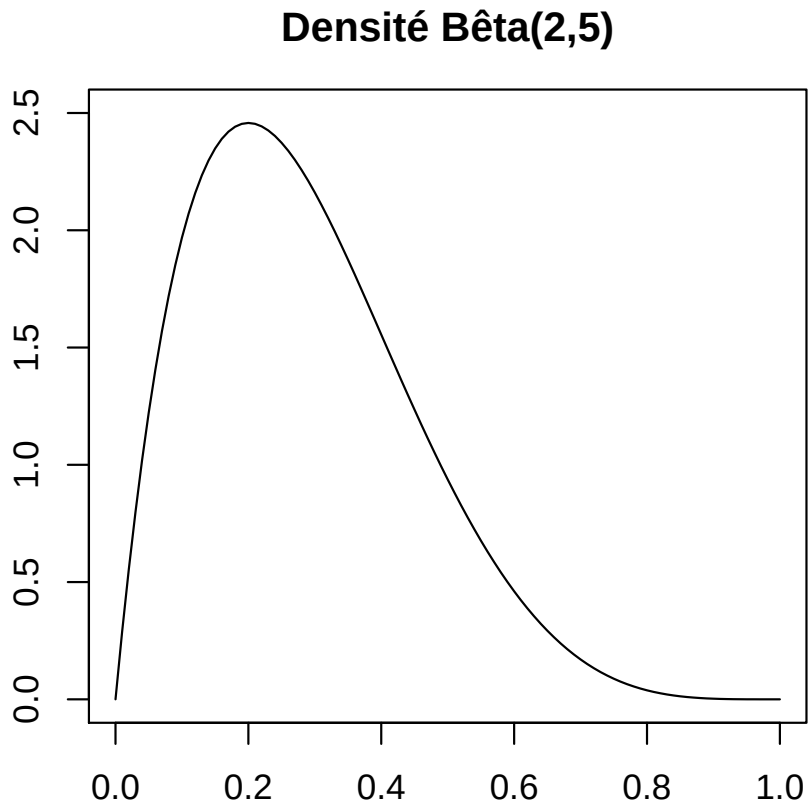
La loi a posteriori satisfait :

$$\pi(\theta|\mathbf{x}) \propto \prod_{i=1}^{n} \left(\theta \frac{\lambda_1^{x_i} e^{-\lambda_1}}{x_i!} + (1-\theta) \frac{\lambda_2^{x_i} e^{-\lambda_2}}{x_i!} \right) \theta^{\alpha-1} (1-\theta)^{\beta-1}$$

Exemple – Simulation d'une loi a posteriori

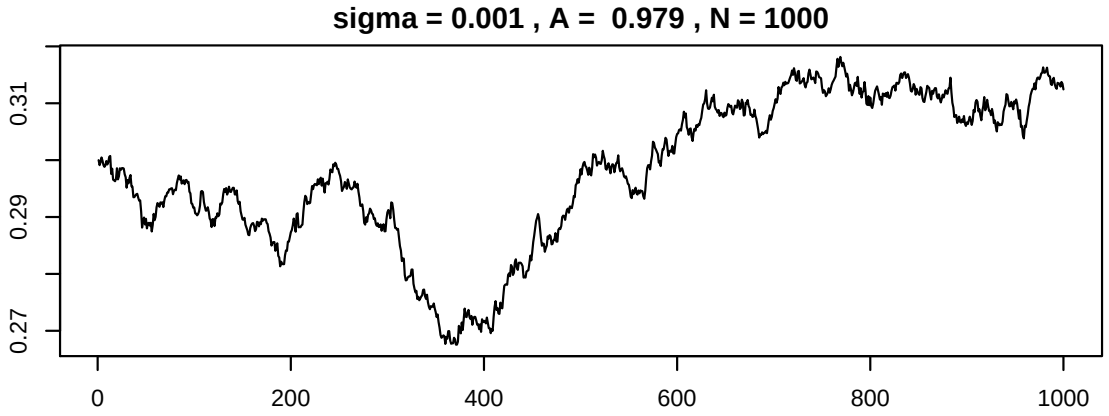
Distributions

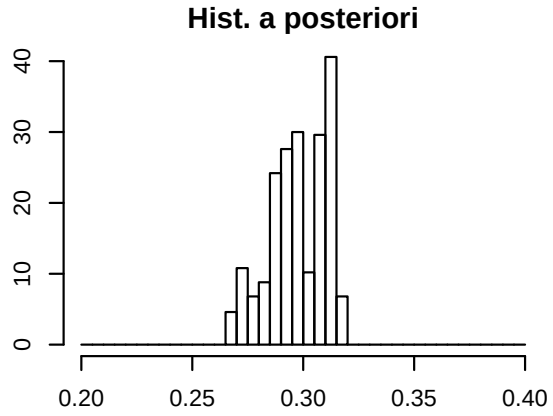


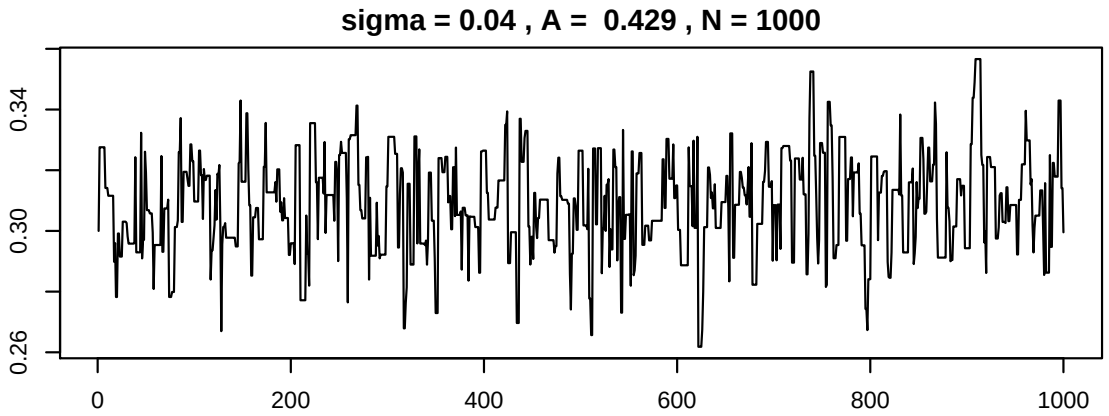


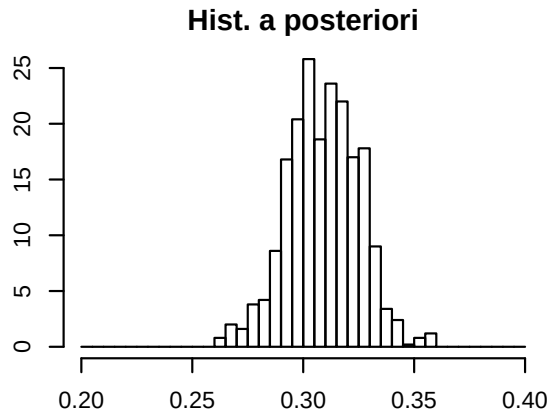
Exemple – Simulation d'une loi a posteriori

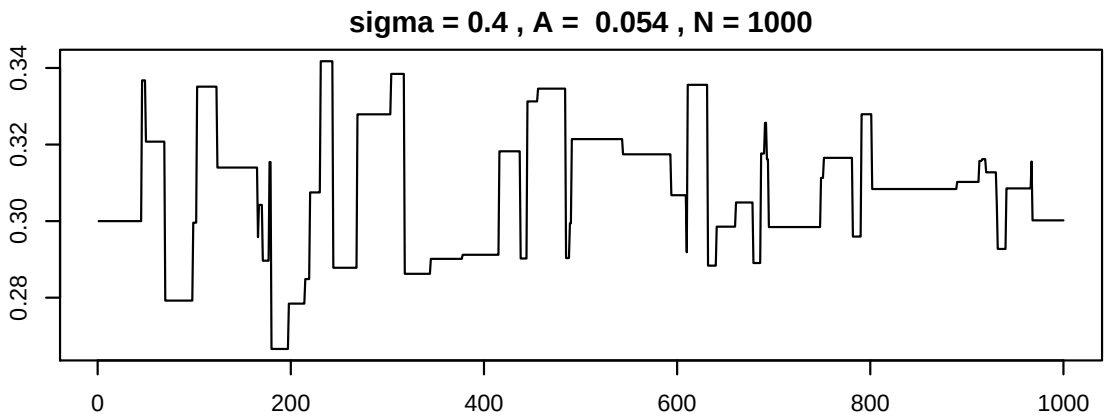
Marche aléatoire Gaussienne : choix de σ

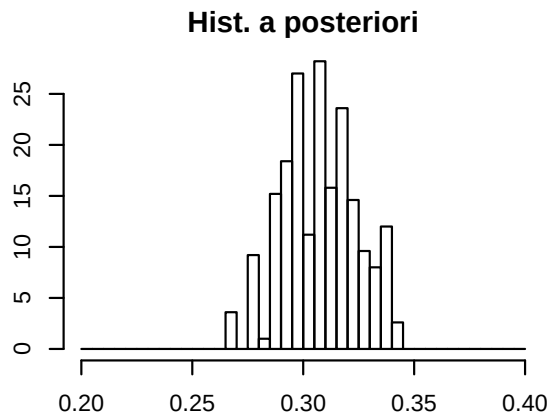






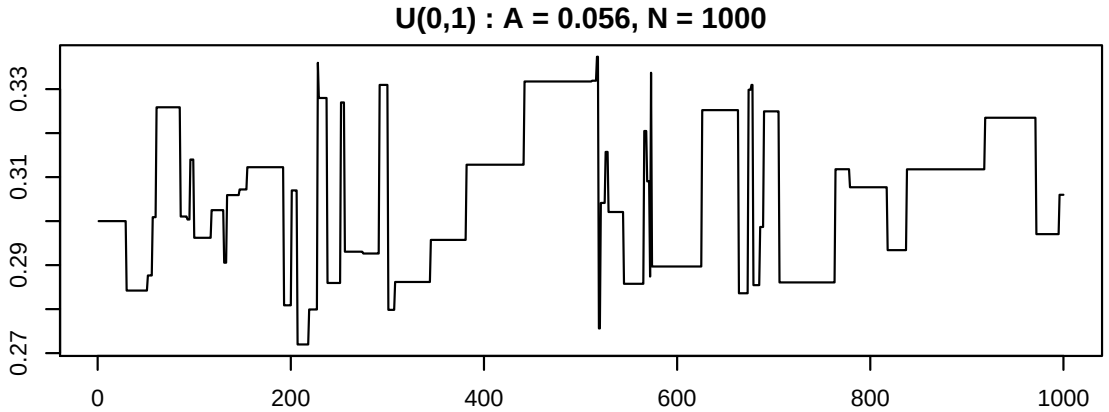


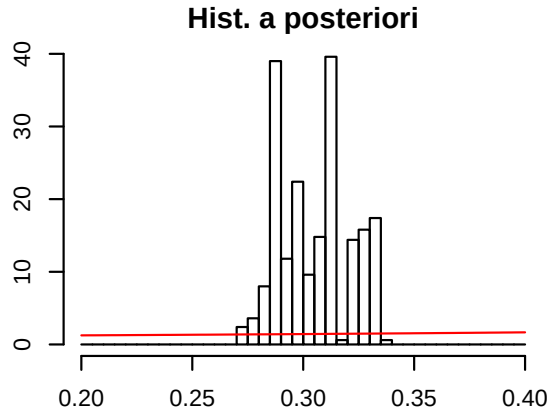


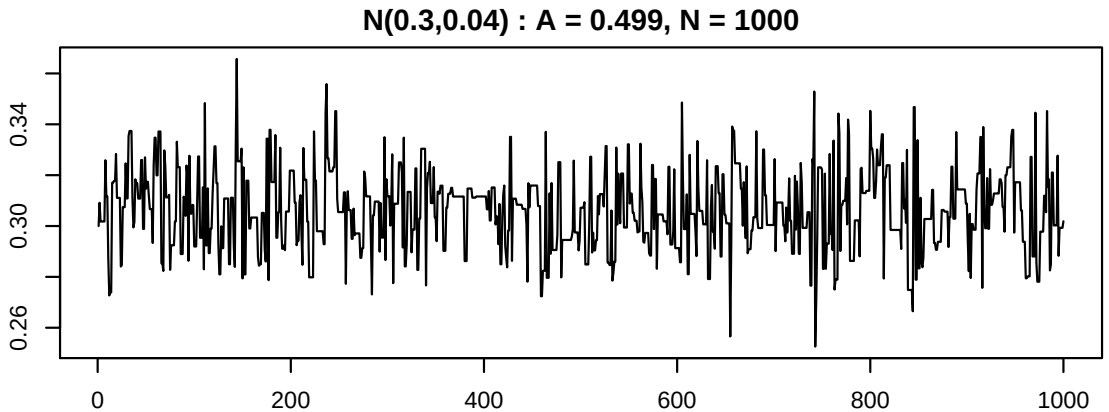


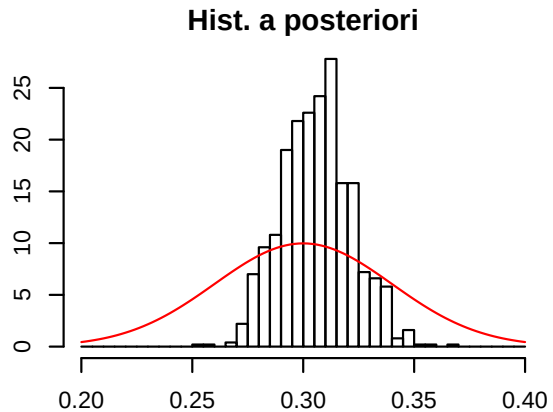
Exemple – Simulation d'une loi a posteriori

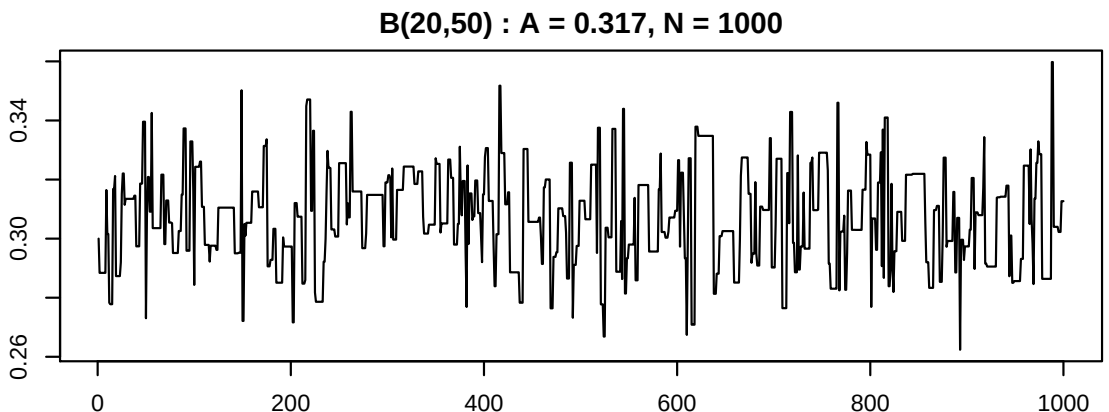
M.-H. Indépendant : choix de la loi instrumentale

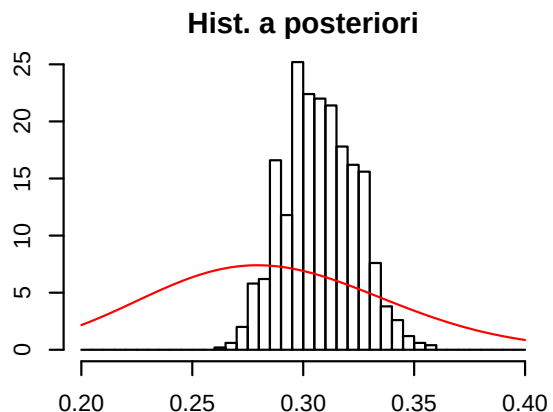












Extensions de l'algorithme

Densité autorégressive

Les candidats $\mathbf{y}$ sont générés à partir d'un processus AR(1), i.e.

$$\mathbf{y} = a - B(\mathbf{x} - a) + \mathbf{z}, \qquad \mathbf{Z} \sim q(.).$$

Par exemple, dans le cas des la normale bivariée, on peut considérer la réflection autour de μ :

$$\mathbf{y} = \mu - (\mathbf{x} - \mu) + \mathbf{z},$$

ce qui constitue une marche aléatoire.

Extensions de l'algorithme

Block-at-a-time

- Lorsque la densité cible est multivariée, on peut effectuer les propositions par blocs;
- Ceci peut être avantageux lorsqu'il est plus facile de trouver des lois instrumentales sur un partie des composantes de la densité cible;
- Peut être fait une variable à la fois;
- En ordre, allée-retour ou permutation aléatoire;
- L'algorithme de Gibbs est un exemple de ce procédé, où le candidat est toujours accepté.

References I

Arora, S. (2013).
Random walks, markov chains, and how to analyse them.

Brooks, S. P. (1998).
Markov chain monte carlo method and its application.
Journal of the Royal Statistical Society, Series D: The Statistician, 47:69–100.

Hastings, W. K. (1970).
Monte carlo sampling methods using markov chains and their applications.
Biometrika, 57:97–109.

Robert, C. P. et Casella, G. (2004).
Monte carlo statistical methods.

References II

Roberts, G. O., Gelman, A., et Gilks, W. R. (1997).
Weak convergence and optimal scaling of random walk metropolis algorithms.
The Annals of Applied Probability, 7(1):110–120.